Journal of Computer and Knowledge Engineering, Vol. 6, No. 1, 2023. (15-26)
DOI: 10.22067/cke.2022.74028.1040

A Taxonomy for RNA Motif Discovery *

Research Article

Zahra Mir[1] Mohammad Allahbakhsh[2] Ali Maghsoudi[3] Haleh Amintoosi[4]

Abstract: Motifs have critical impacts on the behavioral and structural characteristics of RNA sequences. Understanding and predicting the functionalities and interactions of an RNA sequence requires discovering and identifying its motifs. Due to the importance of motif discovery in bioinformatics, a significant corpus of techniques and algorithms have been proposed, each of which has various advantages and limitations and hence, are suitable for specific applications. To understand these techniques and algorithms, compare them, and choose the most suitable one for a particular application scenario, it is crucial to have a clear understanding of the different vital aspects that characterize these algorithms. The lack of such a framework to study these aspects is a serious existing challenge in the literature that needs further investigation. In this paper, we propose a taxonomy and a framework to address this issue. We define the concept of motif discovery process and three aspects that characterize such a process, which are motif type, discovery technique, and application. We then study the literature and classify the existing approaches along with these aspects. This will give the reader a broader view and more precise understanding of what these techniques and algorithms do, how they do it, and what is the most suitable application for each of them. We then present the possible gaps and challenges foreseen to be the future directions of the area.
Keywords: Algorithm, Bioinformatics, Motif Discovery, RNA Motif, Taxonomy

1. Introduction

The fundamental function of RNA is generating proteins through translation. RNA transmits genetic information that is translated by ribosomes into a variety of proteins essential to cellular activities. Three principal types of RNA involved in protein synthesis are messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). RNA is also the principal genetic material for viruses. Moreover, RNA serves other purposes, including gene regulation, RNA interference, and RNA editing. Small regulatory RNAs, which include small nuclear RNA (snRNA), microRNA (miRNA), and small interfering RNA (siRNA), are responsible for these processes [1].

A motif is a small part of a DNA or RNA sequence that has a constant size and is related to DNA and RNA functions. Motifs can represent frequent patterns in biological sequences, such as DNA, RNA, and protein sequences [2]. There is no single description for RNA motifs because they can be presented and analyzed at various levels of RNA structure [3]. RNA motifs can be generally characterized as recurrent structural components that are subject to limitations [4]. In the 1980s, compilations of non-Watson–Crick (noncanonical) pairs were compiled using just the atomic structures of tRNAs and the crystal packing interactions of tiny oligonucleotides [5]. These compilations classified interactions based on the base type (purine-purine, purine-pyrimidine, and pyrimidine-pyrimidine), instead of the geometry [6]. RNA motifs are defined as regulated sequences of non-Watson-Crick base pairs that consist of distinct folds from the phosphodiester chains of RNA strands (Phosphodiester bonds are formed when two hydroxyl groups (OH) in phosphoric acid (H_3PO_4) react with hydroxyl groups (OH) on other molecules to generate two ester bonds. The phosphodiester bond in DNA and RNA is the link between the 3' carbon atom of one sugar molecule and the 5' carbon atom of another, deoxyribose in DNA and ribose in RNA [4, 7]). These sequences might contain motifs at the sequence level that impact the behavior and functionalities of the sequence. At a higher level, RNA molecules fold into distinct secondary and tertiary structures that perform various functional activities (RNA secondary structure is composed of short double-stranded regions separated by tracts of single-stranded nucleotides; tertiary structures are generated by combining the various secondary structure elements [8]). Many of these RNA structures are derived from a set of RNA structural motifs. These structural motifs are repeatedly used in different combinations to form various types of RNA and determine their unique structural and functional properties. Recognizing repeated RNA motifs helps to have a better understanding of RNA structures and the regulatory and functional elements associated with RNA structure [9].

Recent studies have shown that RNA motifs play an essential role in RNA folding. These motifs are often used as nucleation sites for RNA folding. In bioinformatics and biology, motifs are of great importance, and motif discovery has become a very active research field [10]. Motif discovery can be essential for the analysis and understanding of biological information. If a pattern occurs frequently, it must be significant or meaningful. Therefore, the purpose of motif discovery is to extract a variety of biological meanings or concepts from sequences [11]. Moreover, the problem of finding motifs in RNA is a growing issue in computational biology [12, 13]. Many algorithms have been proposed to address the motif discovery problem in RNA. However, due to the lack of a standard criterion, few studies have been performed to compare and classify them.

In this paper, we clarify the problem of motif discovery in RNA sequences. To do so, we first have to have a clear understanding of the process of motif discovery. In a typical motif discovery process, one or more computational techniques are used to discover motifs from a sequence for a

* Manuscript received: 2021 December 6 , Revised, 2022 July 26, Accepted, 2022 December 12.
[1.] Corresponding author. Master of Bioinformatics, Department of Bioinformatics, University of Zabol, Zabol, Iran.
Email: zahramir255@gmail.com,
[2.] Associate Professor of Computer Science, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
[3]. Associate Professor of Animal Science, Department of Animal Science, Tarbiat Modares University, Tehran, Iran.
[4]. Associate Professor of Computer Science, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

specific application. To better understand different aspects of such a process, we ask three key questions and use their answers to build a framework for analyzing motif discovery approaches. The key questions are:

1. What is exactly discovered when we talk about motif discovery? The answer to this question gives us information about the nature and the possible varieties of RNA motifs.
2. How are motifs discovered? The answer to this question gives us a broader knowledge of the various algorithms and techniques that are used for motif discovery.
3. Why do we need to discover motifs? Answering this question leads us to the possible applications for motif discovery techniques.

Relying on the literature review and the answers to the above questions, we propose a taxonomy for motif discovery in RNA. This taxonomy describes the process of motif discovery in RNA and comprises three main aspects, each of which corresponds to one of the above-asked questions. In the first aspect, a general classification of the types of RNA motifs is expressed. In the second aspect, algorithms and methods used to identify and predict motifs in RNA are discussed. Finally, the third aspect describes the application of RNA motifs in different fields. Figure 1 illustrates the proposed taxonomy and the details of the identified aspects. Studying each aspect helps the reader to get a better understanding of the motif discovery process in RNA and offers a comprehensive classification that helps to choose motif discovery algorithms that are more suitable for particular situations.

In the rest of the paper, the different aspects identified and presented in Figure 1 are discussed in sections 2 to 4. Then we look at the state of the art in Section 5. In Section 6, we briefly discuss some remaining open challenges that can be of research interest in the future. We finally conclude the paper in Section 7.

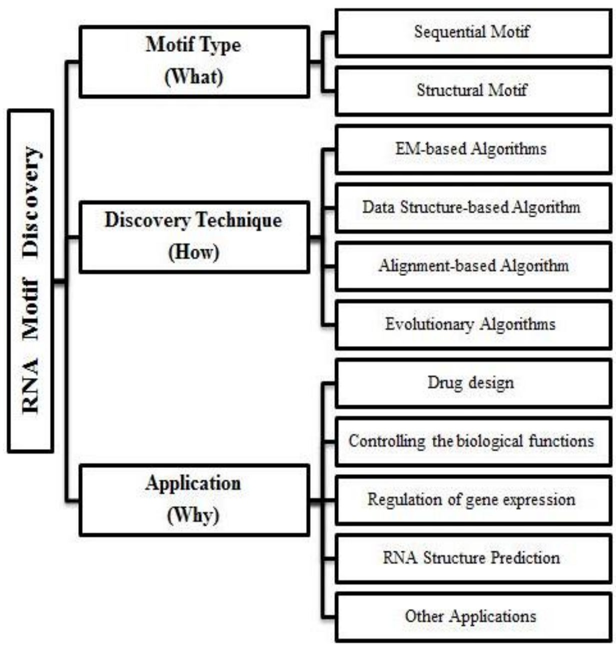


Figure 1. Taxonomy of motif discovery

2. Motif type

Various definitions and terms are used for RNA motifs. Examples of these terms include RNA elements, binding sites, and regulatory elements. RNA motifs can be suggested and analyzed at different levels of RNA structure. Figure 2 illustrates different types of RNA motifs. Generally, the function of motifs in RNA can be summarized as follows: Protein binding, pairing with other RNA, ligand binding site, and correction of a nucleic acid bond [14]. RNA motifs can be classified into different categories in the motif type aspect. For example, one of these categories may include the following categories: Sequential motif and structural motif.

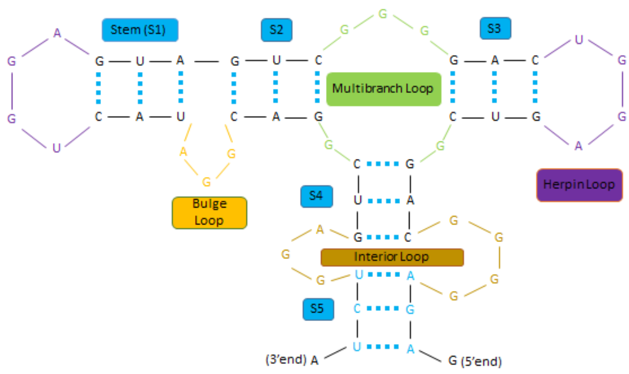


Figure 2. RNA secondary structures with basic structural motifs

2.1. Sequential motifs

Sequential motifs are frequentative patterns in biological sequences, and can work at the RNA sequence level [3].

2.2. Structural motifs

An RNA structural motif is characterized as a set of nucleotide sites that fold into a stable three-dimensional structure and is abundantly found in natural RNAs. Moreover, a structural motif may have cofactors such as water molecules, metals, or other ions to protect its spatial structure [15]. Structural motifs are supersecondary structures in the RNA structure that can have multiple biological actions and play functional and structural roles. In this section, we focus on RNA structural motifs and define the types of known structural motifs in RNA.

1. Hairpin loops: The hairpin loop is a crucial element and is practical in several known RNA systems. Hairpin loops connect the 3'-and 5'-ends of a double helix. These loops are formed when the RNA strands fold together [16]. Examples of hairpin loop motifs are the T-loop [17], D-loop motifs of tRNA [16], the lonepair triloop [18], and the sarcin-ricin loop [19].
2. Internal loops: An internal loop detaches double helical RNA into two fragments by positions that are not Watson–Crick paired in at least one strand of the double. Sometimes interpolation on just one strand is specified as "bulge loops" [15]. Some RNA internal loop motifs include the kink-turn (K-turn) [20] reverse kink-turn [21] and hook-turn [22].
3. Junction loops/multiloops: Junctions refer to areas with two or more double-stranded stems within a molecule and adjacent together. Junction loops are organized by the intersection of three or more double helices. Double helices are segregated by single-strand sequences. N-

linker sequences are N helices connectors in a junction loop. Examples of junction loops are tRNA and hammerhead ribozyme [15].

4. Helices: The RNA molecule composes Watson-Crick base pairs with A-form double-stranded helices. Nevertheless, in different environmental conditions, including high salt concentration, different forms of RNA helices have also been observed [23].
5. RNA bulges: Bulges are unpaired nucleotides that are enclosed in a strand of double-stranded nucleic acids formed by hydrogen bonds, including the usual Watson-Crick bonds and unusual base pairs. The bulge sizes vary from an unpaired nucleotide to several nucleotides [24].
6. Bining motifs: One of the basic functions of RNA is to act as binding ligands. They also play the role of cofactors, substrates, and signals for structural stability. Binding motifs are important for riboswitch, ribozyme, and splicing functions, as well as in interceding RNA–protein, RNA–RNA intermolecular, and tertiary interactions [15].
7. Metal binding: Examples of metal ion interactions with RNA include diffusion ions that compress near RNA because of the electrostatic field while maintaining their hydration sphere and chelating ions that are in direct contact with RNA and act as a specific site that may replace some of the water molecules around the metals with the polar RNA atoms [25].
8. Natural and selected aptamers: Aptamers are oligonucleotide or peptide molecules that bind to a specific target molecule. Natural aptamers are present in riboswitches. At least three classes of RNA can bind specifically to small ligands, including functional RNAs, the riboswitches found in the untranslated regions of mRNA and the in vitro selected (SELEX) aptamers [15].
9. Tertiary interactions: RNA molecules are stacked into the exact tertiary structure to form spherical shapes that are stabilized by a large number of interactions (RNA molecules with tertiary folding). Third-fold RNA molecules are identified by ligands, other RNA molecules, and proteins, which lead to biochemical reactions that affect cellular metabolism. The tertiary structure of RNA biomolecules, like their secondary structure, is created by a limited number of repetitive motifs and interactions [26]. The tertiary structure is formed when the secondary structures are formed between van der Waals interactions. Generally, tertiary interaction structures include two helices, two unpaired and one unpaired area, and one standard double helix. These tertiary structures and motifs include the following cases.
10. Coaxial helices: By overlapping, nucleotide bases are formed by two independent helices, coaxial helices, or pseudo-continuous. If two helices are close together, their ending bases are stacked together and the helices line up and form a coaxial [27, 28].
11. Kissing loops: Kissing loops are the long range of the base pairing between the stem and hairpin loops, and their interactions are accomplished by loop nucleotides that complement Watson-Crick's interaction. Kissing loops are found in a variety of RNAs including rRNA, tRNA, and mRNA [29, 30].
12. Ribose zipper: The ribose zipper is organized by hydrogen bonding between backbone ribose 2' hydroxyls from two distant regions of the chain. Ribose zipper secondary structures include double helix, Hairpin, or internal loop [28, 31].
13. A-minor motif: A bunching of A-minor interactions is often declining in type and order to go from the 5'to the 3'direction. A-minor motif secondary structures include the internal loop and hairpin loop [32].
14. D-loop and T-Loop: T-loop and D-loop motifs are created when a complex interaction occurs between two conserved hairpins in tRNA, including paired bases. The secondary structure is the hairpin loop [33, 34].
15. Tetraloop receptor motif: The tetraloop receptor motif is one of the most common motifs among the RNA tertiary interactions. This motif is involved in the crystal structure of large RNAs and the construction of RNA nanostructures.
16. Kissing hairpin loop: When Watson-Crick bases pair between the single-stranded residues of two hairpin loops with complementary sequences, they form a kissing hairpin [29, 31]. These structures play a role in controlling biological functions using the antisense phenomenon.
17. Pseudoknot: One of the most complex tertiary structures in RNA is Pseudoknot, one of the most important features of large RNA structures. Pseudoknot is formed by base pairing between nucleotide loops (hairpin or internal) and bases outside the enclosing loop [35, 36]. Pseudoknots have a very diverse structure and can be involved in some viruses in regulating gene expression by translocating the ribosomal translation template.

3. Discovery technique

The motif finding problem can be crucial to the analysis and comprehension of biological information [11]. Many algorithms have been implemented to solve the motif discovery problem, while each of these algorithms identifies different types of motifs.

In the technique aspect, we discuss a number of algorithms used to identify and predict motifs in RNA. Algorithms can be classified into different categories in the technique aspect. For example, one of these categories may include the following categories: EM-based, Data structure-based, alignment-based, and evolutionary algorithms.

3.1. EM-based algorithm

The algorithms in this group use the EM (expectation maximization) algorithm to extract and discover the motif. EM algorithm is an iterative algorithm that seeks to find maximum likelihood estimates of unknown parameters. This algorithm is a standard method when random variables are hidden.

EM algorithm consists of two steps: The first step is known as the "Expectation" step or E-step. This step involves computing the expected value of the hidden variables, and gives the current parameter values. The next step is known as the "Maximization" step or M-step. This step calculates a new set of parameters by maximizing an appropriate function that gives the expected value of the hidden variables. Some of the advantages of EM algorithms

are: E-step and M-step are often easy for many problems in terms of implementation, EM algorithm always guarantees that likelihood will increase with each iteration, and solutions to the M-steps often exist in the closed form. Moreover, some of the disadvantages of EM algorithms include that they have slow convergence, they make convergence to the local optimal only; and they need both the probabilities, forward and backward (numerical optimization requires only forward probability). The following examples refer to some EM-based algorithms:

1. The MEME algorithm simultaneously identifies the motif and the position of the motif in a set of sequences. The motif is designed as a special position weight matrix, where in each motif location there is a probability of the residue or nucleotide. MEME uses an EM algorithm for learning, which in the first step calculates the position weight matrix values, and in the second step, it obtains the expected value of the hidden variables. As a result, the hidden variable with the highest expected value in a given sequence specifies an estimate of the motif location. The goal of MEME algorithm is to discover new motifs in a set of biopolymer sequences, when there is little or no prior knowledge of possible motifs [37, 38].
2. CMfinder algorithm is based on the expectation maximization (EM) algorithm and the covariance model. The algorithm starts with finding possible motif patterns and looks for motifs that have a more stable structure. For each input sequence, the minimum free energy from all subsequences is calculated. Then the best patterns are selected and compared at the basic level. The algorithm is obtained by constructing a primary alignment of similar patterns. The primary alignment is used as a seed for EM algorithm. Afterward, EM algorithm is performed twice: First, to correct the primary alignment, and second, to scan the sequences. Ultimately, the motifs are combined to achieve the final result [39].
3. RNApromo is one of the methods based on the EM algorithm for identifying RNA motifs, which starts with a set of RNA sequences and a set of proposed secondary structures. The motif prediction algorithm consists of two main parts. In the first part, various heuristic methods are used to identify the motif candidate. In the second part, the candidates are corrected using EM algorithm [40].
4. MEMERIS algorithm is based on MEME algorithm and is a method for finding motifs in single-stranded regions. The algorithm attempts to detect specific sequence motifs associated with binding sites for mRNA proteins. These sites are often located in the 3' UTR or 5' UTR of mRNA and mostly in specific standard areas [41].

3.2. Data structure-based algorithm

Data mining consists of many data analysis techniques whose primary purpose is to discover the hidden and useful data model of a pervasive data set. A data structure is a data organization, management, and storage format that can be used effectively. Some examples of data structures are: Array, linked list, tree, graph, stack, and queue. Usually, efficient data structures are important for designing efficient algorithms. As an outstanding example, in recent years, data mining has seen extensive development in graph mining [42]. Graph mining algorithms are helpful in discovering and analyzing graph data. Graph-based algorithms are generally determined in terms of the number of vertices (|V|) and the number of edges (|E|) in the input graph. In molecular biology, each vertex corresponds to a biomolecular entity, and each edge represents a specific connection or interaction between such entities. By knowing the RNA secondary structure, we can turn the problem of finding common patterns among a set of RNA sequences into a problem of exploring repeated tree patterns in a set of trees (secondary structures of the RNA molecule) [43]. Some of the advantages of graph-based algorithms are: 1) Graph-based algorithms may not always be the most accurate, but they are generally very efficient; 2) Graph-based algorithms are generally easier to write as well as explain. Moreover, one of the disadvantages of graph-based algorithms is that they do not always lead to the most optimal path. Motif discovery algorithms can use data structures to speed up access and retrieve words. The following examples refer to some data structure-based algorithms:

1. Structural relation matching is a technique for identifying structural patterns in the secondary structure of RNAs and RNA-RNA interactions or their abstractions (cores and shadows) that focuses on patterns and structures with arbitrary pseudoknots [44, 45]. This method introduces four algorithms: Loop determination, core determination, determination of the relation matrix, and structural relation matching. The loop determination and core determination algorithms receive the secondary structure of an RNA or RNA-RNA interaction as input and generate a list of secondary structure loops, core, and shadow, respectively. The relational matrix is determined using an algorithm that maps each RNA secondary structure into a matrix. Finally, the structural relation matching algorithm searches for a submatrix to identify each pattern of the RNA structure [46].
2. Mauri and Pavesi developed an algorithm that uses affix tree data structure to discover internal loops, bulges, and hairpin motifs in RNA. Substrings with definite lengths appearing in at least *q* sequences are found and expanded [47].
3. Seed algorithm uses a data structure called suffix arrays to discover RNA structural motifs. The algorithm uses data structures to store the sequence, its reverse, and the input sequences [48].
4. IncMD, as an algorithm for identifying RNA secondary structural motifs, works incrementally. IncMD employs data structures such as trie-based linked lists of prefixes (LLP) to speed up pattern searching and retrieval, supporting count, and candidate development. Using nesting and joining processes, IncMD generates frequent patterns incrementally from RNA secondary structure basic items. Nesting enables the addition of a base pair, an internal loop or a bulge to an available structure to develop new structures. The joining procedure employs adjacency information between two structures to produce a new structure [49].
5. ComRNA is a graph-based approach that can find a structural motif called a pseudoknot. ComRNA uses an n−partite undirected weighted connector graph and

seeks to find all possible stable stems of length L. In this graph, the vertices represent the stems, and the edges indicate the similarity between them. Then, a graph method like topological sorting is applied to find the best sets of stems [50].

6. GraphClust2 is a scalable clustering method for RNAs based on sequence and structural similarity. The GraphClust method incorporates sequence and structure information into high-dimensional sparse feature vectors by using a graph kernel technique. A locality-sensitive hashing approach is also used to efficiently cluster these vectors, with a linear time complexity over the number of sequences. This method offers a comprehensive solution that can start with raw high-throughput sequencing and genomic data and conclude with predicted motifs with comprehensive visuals and metrics [51].
7. RNAmine algorithm uses graph mining methods to discover stems in a set of RNA sequences. A directed labeled graph is created, called the stem graph, which represents the RNA sequence and its secondary structure. In this graph, the vertices represent the stems, and the edges indicate the relative positions of two stems. RNAmine algorithm uses a branch and bound algorithm to calculate the stem pattern [52].

3.3. Alignments-based algorithm

Multiple sequence alignments (MSA) are among the fundamental techniques for biological sequence analysis, such as DNA, RNA, and protein sequences in bioinformatics and computational biology. There are many different approaches for generating multiple sequence alignment, and their use depends on sequence length, type, and the user precedences. Some examples of multiple sequence alignment algorithms are BLAST, PSI-BLAST, ProbCons, T-Coffee, MUSCLE, MAFFT and ClustalW. Some of the advantages of alignments-based algorithms are: Improve accuracy, they are fast, and output alignments generally are good within clusters of closely related sequences. In addition, some of disadvantages of alignments-based algorithms are: 1) Weak scalability; 2) Some algorithms cannot fix errors; and 3) Some algorithms are computationally expensive (slow problem on large inputs). The following examples refer to some alignments-based algorithms:

1. Foldalign is a dynamic programming-based algorithm that is used to discover motifs for two sequences. It is based on the Sankoff's algorithm [53]. Foldalign maximizes number of base pairs and alignment resemblance in two aligned sequences [54].
2. Sankoff algorithm is a dynamic programming-based algorithm and seeks to simultaneously align and predict the structures of two or more RNA sequences in a global way. The purpose of Sankoff algorithm is to minimize the summation of the alignment distance and the free energy of the involved RNA molecules [53].
3. MCCaskill is an efficient dynamic programming algorithm that simultaneously calculates the alignment and secondary structure [55].

3.4. Evolutionary algorithms

EA is an algorithm that combines dimensions of natural selection and continuity of coordination. Evolutionary computational algorithms search for optimal solutions according to the concepts of biological evolution, such as reproduction, natural selection, and survival of the fittest [56]. Evolutionary algorithms include significant components such as selection, reproduction, fitness function, solution representation, and population initialization. Some of the advantages of evolutionary algorithms are: 1) Evolutionary algorithms are able to solve problems that have unknown solutions; 2) EA algorithms can be used to robust the dynamic changes; and 3) Evolutionary algorithms have ability for self-optimization. Moreover, one of the disadvantages of evolutionary algorithms is that there is no guarantee that the algorithm will converge to a global optimum. The following examples refer to some evolutionary computational algorithms:

1. RNAGA seeks to discover common secondary structures in a set of homologous RNA sequences using a genetic algorithm (GA). Genetic algorithms are based on the ideas of natural selection and genetics and in RNAGA, it is applied at different levels. In the first stage genetic algorithm is applied to each sequence to obtain a set of stable structures. Afterward, it is applied to the set of stable structures [57].
2. GPRM algorithm uses genetic programming (GP) to discover structural motifs in RNA sequences. The algorithm needs two sets of inputs that include positive and negative sets. A positive set includes a set of co-regulated RNAs, and a negative set is a set of randomly generated sequences [58].
3. GeRNAMo uses genetic programming (GP) to discover motifs in RNA and begins with a set of suboptimal secondary structures. It uses genetic programming on the output of RNAsubopt [59].

4. Application

RNA motifs can have a variety of basic applications in various fields. For example, RNA motifs can be used in basic applications such as drug design, controlling biological function, regulation of gene expression in some viruses, and manufacturing of RNA nanostructures. In this aspect, we aim to examine the application of some different types of RNA motif discovery.

4.1. Drug design

Drug design is the process of discovering new drugs based on a specific biological target. The drug is usually a small organic molecule which activates or inhibits biomolecular functions [60]. For example, Remdesivir is one of the drugs offered to fight the viral infection caused by SARS-CoV-2. The drug works by targeting RNA-dependent RNA polymerase (RdRp), which is essential for the replication of the SARS-CoV-2 virus. The RNA-dependent RNA polymerase (RdRp) is made of three main nsp (non-structural proteins) called nsp7, nsp8, and nsp12. Among these nsps, nsp12 RdRp, which has seven conserved motifs from A to G in its structure, plays a more important role in RNA-dependent RNA polymerase (RdRp) function. For RNA-dependent RNA polymerase (RdRp) inhibition, Remdesivir exerts a more detailed and precise impact on these motifs. Eventually, this effect will stop the virus replication [61]. As another example of applications, in some cases, loops can have medicinal purposes. For example, in

TAR regions of the HIV virus genome, there is a loop-stem structure in which three nucleotides are unpaired. This sequence binds to the arginine amide protein group or a 37 amino acid peptide called ADP-1. As a result, the trans-activator of transcription protein cannot bind to the TAR and ultimately does not transcribe [62].

The stem-loop II motif possesses the features of an important prospective therapeutic target across coronaviruses and astroviruses that contain it, due to its high sequence conservation, complicated tertiary structure, and cell dependant structural variations [63]. Studies indicate that due to a G to U base transversion, the stem-loop II motif in SARS-CoV2 has a various conformational profile than that of SARS-CoV, therefore this conformational profile is modified upon ligand binding [64]. As a result, considering that the conformation of noncoding RNA components generally dictates their function [65], the coronavirus stem-loop II motif could be a druggable target.

4.2. Controlling the biological functions

In general, biological functions are mechanisms that take place in the bodies of living organisms. Kissing hairpin is one of the structures that are involved in controlling biological functions using the antisense phenomenon.

4.3. Regulation of gene expression

In order to show gene effects, mechanisms must be put in place to produce the specific product of that gene (protein or RNA) and to reduce or increase the amount of that product when needed. These mechanisms in the cell are called gene expression regulation. Depending on how the processes are evaluated, activities associated with gene expression can operate at various times during transcription or translation [66]. For example, pseudoknots have a wide variety of structures and can be involved in some viruses in regulating gene expression by displacing the ribosomal translation format.

4.4. RNA structure prediction

According to structural and comparative studies, the structure of RNA is largely modular and formed of motifs or repeating building blocks. These patterns are visible at the primary, secondary, and tertiary structural levels. Examples of secondary structure motifs that define hydrogen-bonding patterns include hairpins, junctions, internal loops, and stems. Moreover, RNA tertiary (3D) motifs are structural elements that occur repeatedly and are subject to various RNA-RNA interaction constraints [67]. RNA 3D motifs play a pivotal role in the folding of RNA as well as biochemical functions. The determination of RNA 3D motifs not only enhances our knowledge of the principles of organization of complex RNA structures but also serves as a starting point for applications to synthetic biology and nanodesign [68-70].

4.5. Other applications of RNA motifs

The identification of patterns in DNA, RNA, and protein sequences has resulted in the solution of many important biological challenges. Finding regulatory information within either DNA or RNA sequences and protein domains, RNA transcription, identifying binding sites in DNA, RNA, or protein, determining open reading frames, identifying promoter elements of genes, locating RNA degradation signals, identifying intron/exon splicing sites, and finding RNA degradation signals are just a few examples [71].

Kissing loops are found in a variety of RNAs, including tRNA, mRNA, and rRNA. Kissing loops are commonly used in retroviruses to initiate the dimerization process in genomic RNA. Tetraloop-receptor motifs are involved in the crystal structure of large RNAs and the construction of RNA nanostructures. Binding motifs are important for the function of ribosomes and riboswitch, splicing, as well as the internal molecules of protein-RNA, RNA-RNA, and the tertiary interaction. The sarcin-ricin motif is one of the motifs found in prokaryotes, especially bacteria, and some eukaryotes. This motif is heavily preserved in the large ribosomal RNA subunits of 28 s and 23s. The sarcin-ricin motif is a suitable place to connect EF-Tu and EF-G lengthening factors [19].

5. State of the art

In this section, we present a summary of the state of the art in RNA motif discovery. We first look at the sequence databases that are available online and are used for both research and development purposes. We then introduce a set of available tools that are used as motif discovery tools and software. Finally, we study a selected set of related research works, and analyze them, according to the framework and the aspects we proposed in this paper.

5.1. Databases

In addition to the sequences that are identified and used in research labs all around the world, there are databases that contain registered and approved RNA sequences. These sequences are used in research activities as baselines, testbeds, or case studies. They can also be used as a means for testing and approving the performance of software tools that are related to analyzing sequences.

These databases, also known as biological databases, can be divided into three categories based on their content: primary, secondary, and specialized databases [72].

The primary databases include the original row of biological data. There are three primary sequence databases that store the raw sequence data of nucleic acids produced and sent by researchers from around the globe. These databases include GenBank, EMBL (the European Molecular Biology Laboratory), and DNA Data Bank of Japan (DDBJ).

The secondary databases contain data that is derived from analyzing the data in primary databases. These databases are usually of higher accuracy levels. They combine manual and automated methods to produce new knowledge out of the existing data sources

There are also databases that store information about a particular research topic or focus on a specific organism. These databases, also called specialized databases, can contain sequences or other forms of information. Many genome databases, such as Flybase, Wormbase, AceDB, and TAIR, fall into this category [72].

Table 1 illustrates some of the major biological databases that are available online.

Table 1. Some data sources available on the internet

Databases and Retrieval Systems	Brief Summary of Content	URL
GenBank	Primary nucleotide sequence database in NCBI	www.ncbi.nlm.nih.gov/Genbank
EMBL	Primary nucleotide sequence database in Europe	www.ebi.ac.uk/embl/index.html
DDBJ	Primary nucleotide sequence database in Japan	www.ddbj.nig.ac.jp
Entrez	NCBI portal for a variety of biological databases	www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
Rfam	Provide alignments, consensus secondary structure and covariance models for RNA families.	http://rfam.org
Ribosomal database project	Ribosomal RNA sequences and phylogenetic trees derived from the sequences	http://rdp.cme.msu.edu/html
RSMDB	RNA Structural Motif Database	http://www.rna.ucf.edu/RNAMotifWebsite/index.php
RNAStructuromeDB	A genome-wide database for RNA structural inference	https://structurome.bb.iastate.edu
RNA CoSSMos 2.0	A database of secondary structure motifs in RNA three-dimensional structures.	http://rnacossmos.com
RBPDB	The database of RNA-binding protein specificities	http://rbpdb.ccbr.utoronto.ca

5.2. Discovery tools

Due to the extensive spectrum of applications, active research groups, and businesses that are active in the area of RNA motif discovery, there exist a large number of tools that are used for motif discovery. As a result of our literature research and checking the existing tools, we selected a set of tools that are more popular and applied in the area.

1. BML: Bipartite Motifs Learning (BML) is a web tool that uses high-throughput sequencing data as input to give a user-friendly gateway for online exploration and study of sequence motifs. BML employs two approaches to find motifs: PWM (Position Weight Matrix) and DWM (Dinucleotide Weight Matrix). It also offers two modes for finding motifs: with parameters and parameters-free (PF). Users can specify default settings or a small set of input parameters in the parameter mode, including the type of sequence (DNA or RNA), forward or reverse orientation, whether or not to enable degenerate motif sites, motif lengths and spacing, predicted motif site distribution in the sequences, and the number of iterations to execute the algorithm [73].
2. MDS2 is a motif discovery tool based on short nucleotide sequences. In MDS2, the motif search starts with finding over-displayed di-mers in the given sequences and then seeks longer motifs by finding a significant path in the graphs. Motif detection in MDS2 involves several steps, including construction of a di-mer graph (each node indicates a di-nucleotide, and each directed edge links two overlapping di-nucleotides), k-mer (substrings of length k) [74] search and evaluation, construction of each k-mer similarity graph, motif candidate identification, motif clustering, and final motif selection [75].
3. FR3D: Finding recurrent three-dimensional (3D) motifs in RNA (FR3D) is one of the application software in determining the three-dimensional structure of RNA. Moreover, with the help of this software, you can search for specific types of motifs from RNA such as hairpin loops and junctions in different organisms from bacteria and viruses to higher organisms [76].
4. RNA Structure: The software is designed to model the structure of DNA and RNA, compare base pairs and modified base pairs. RNA Structure is based on differences in the free energy of structures, and the accuracy of the software shows that sub-structures such as pseudoknot can also be modeled [77].
5. MoD: MoD (Motif Discovery) Tools online server is a series of tools for finding novel conserved sequence and structural motifs in nucleotide sequences, which are then potential candidates for regulatory activity [78]. The following programs are available on the server: RNAprofile [79], which searches for motifs in non-coding RNAs that are conserved in both sequence and secondary structure; Weeder [80], for the finding of conserved motifs in nucleotide sequences from co-regulated genes; and WeederH for discovering motifs and distal motif modules in homologous gene sequences.
6. BEAM: BEAM is a web server that can identify RNA secondary structure motifs and work with tens of thousands of molecules. The server is designed to simplify data preprocessing by automatically folding and encrypting RNA sequences, giving users an option for the preferred folding program [81].
7. BRIO web server: BRIO (BEAM RNA Interaction Motifs) is a new web service for identifying sequence and structural RNA-binding motifs in one or more RNA molecules [82]. BRIO uses BEAR (Brand nEw Alphabet for RNA) [83, 84], encoding to describe structural motifs. The structural context of each nucleotide can be included in the secondary structure representation using this string encoding without adding algorithm complexity.
8. MEME: MEME service is used to perform a wide range of motif-based analyses and includes various tools for discovering new motifs in a set of nucleic acid and protein sequences. Some examples of MEME tools are FIMO, MAST, GLAM2Scan, MCAST, Tomtom, SpaMo, CentriMo, MEME-ChIP, and GOMo [38].
9. SSMART: SSMART (Sequence-structure Motif Identification) is a de novo motif finder that uses

massive sets of RNA sequences collected from genome-wide in vivo or in vitro experiments like CLIP [85, 86, 87], or RNAcompete [88] to identify sequence-structure binding motifs. SSMART generates easy-to-understand sequence-structure binding motifs by modelling the primary sequence and structural features of RNA target sites at the same time. This technique looks for optimum sequence-structure motifs of variable length in potential RBP binding sites that are prioritized by experimentally obtained binding evidence [89].

10. SPOT-RNA: SPOT-RNA is the first predictor of the RNA secondary structure that can predict a variety of pairs (canonical, non-canonical, pseudoknots, and base triplets) [90].
11. VS fold/VS Subopt: It predicts the RNA secondary structure and pseudoknots using an entropy model derived from polymer physics. The vs_subopt program calculates suboptimal structures based on free energy from vsfold5 [91].
12. GraphProt: GraphProt is a graph kernel-based machine learning technique that extracts binding motifs from a set of bound and unbound sequences. The main goal of GraphProt framework is to use cross-linking immunoprecipitation-high-throughput sequencing (CLIP-seq) data to learn binding preferences, then use trained models to find motifs of sequence and structure binding preferences and forecast new RNA-binding protein target locations within the same organism [92].
13. CyloFold: CyloFold is a method to predict the secondary structure based on helices that accept complex pseudoknots [93].
14. STREME: STREME is a flexible tool for discovering sequence motifs in a huge data collection of sequences. The sequences may be in the DNA, RNA, or protein alphabet, or in a custom alphabet. STREME can detect motifs in datasets with hundreds of thousands of sequences, as well as short and long motifs (from 3 to 30 locations). STREME also can find differential motifs in pairings of sequencing datasets, analyses, and reports the statistical importance of each motif it finds [94].
15. RNAmine: A software tool for extracting structural motifs from a set of RNA sequences, for example, finding frequent stem patterns from a set of unaligned RNA sequences [52].
16. TFold: TFold is a tool for predicting non-coding RNA secondary structures, including pseudoknots. This tool takes the aligned RNA sequences as input and returns the predicted secondary structures. It also combines stability, conservation, and covariation criteria to search for stems and pseudoknots [95].

5.3. Research prototypes

Based on our literature review, a big corpus of research is proposed in the area of motif discovery. Due to the interdisciplinary nature of bioinformatics, the related works belong to a very broad spectrum, ranging from mathematical models and data mining techniques to biological research works that are based on lab experiments. Some of these researches are referred to in some parts of this paper. We shortlisted the related literature and came up with a list of some related works that are more popular in this field. Table 2 shows a list of eleven selected related works. The selected works are not meant to cover all the literature. However, we use this table to show how the related works can be analyzed, classified, and compared using the proposed framework and identified aspects.

Table 2. List of selected related works

Article	**Motif Type Aspect**		**Technic Aspect**				**Application**
	Sequential motifs	**Structural motifs**	**EM-based algorithm**	**Alignment-based algorithm**	**Data structure-based algorithm**	**Evolutionary algorithm**	
[38]	✓		✓		✓		Biological functions, regulation activities, drug design
[39]		✓	✓	✓			Biological activities
[47]		✓			✓		Regulation of gene expression
[48]	✓	✓			✓		Regulation of gene expression
[50]		✓			✓		Regulation of gene expression
[52]		✓			✓		Biological activities
[54]	✓	✓		✓			Biological activities
[58]		✓				✓	Regulation of gene expression
[59]		✓				✓	Regulation of gene expression
[75]	✓				✓		Understanding the mechanism of miRNA loading and intercellular transfer
[96]	✓			✓	✓		Regulation of gene expression

Bailey et al. [38] introduced the MEME Suite web server, which offers a unified platform for the online finding and analysis of sequence motifs that describe characteristics, including DNA binding sites and protein interaction domains. This software uses an EM-based algorithm, with the following applications: Biological and regulatory functions and drug development. Yao et al. [39] proposed the CMfinder tool that predicts RNA motifs in unaligned sequences. It is an expectation maximization algorithm that uses covariance models to describe motifs and generates a motif structural alignment and statistical model that can be used directly for homology search. CMfinder is an effective tool for discovering novel noncoding RNAs (ncRNAs) and a number of other biological activities. A data structure-based algorithm for pattern matching and discovery in RNA secondary structures was developed by Mauri and Pavesi [47].

Various functional or regulatory roles are carried out by some of these structures. Anwar et al. [48] considered a data structure-based algorithm called Seed that searches the space of RNA sequences and structural motifs exhaustively, thereby aiding to the identification and characterization of consensus structures. These discoveries play an important role in the regulation of gene expression. Ji et al. [50] used comRNA, which employs a graph-theoretic technique to predict common RNA secondary structure motifs, such as pseudoknots in unaligned sequences, which play important roles in the transcriptional and post-transcriptional regulation of gene expression. Another algorithm is RNAmine [52], which extends graph mining techniques to extract stem patterns from RNA sequences. These structures serve crucial roles at various biological stages. FOLDALIGN, proposed in [54], is an alignment-based technique that identifies the most significant common sequence and structure motifs in a collection of RNA sequences. This technique can be used to discover regulatory regions and various biological functions. Hu [58] proposed an evolutionary algorithm that can find structural motifs more complicated than stem-loop structures. Understanding the structural motifs can help us acquire a deeper insight into regulating activities. The GeRNAMo (Genetic Programming of RNA Motifs) proposed in [59] employs an evolutionary computation system that seeks a motif shared by a given set of RNA sequences. Discovering these structures is helpful in regulating activities. Gao et al. [75] considered a method for motif discovery based on short nucleotide sequences (e.g., small non-coding RNAs or truncated messenger RNAs). This method uses a data structure-based algorithm (such as the graph algorithms). One of its key applications is understanding the mechanisms of miRNA loading and intercellular transfer. Chen et al. [96] demonstrated that LocalSTAR3D is capable of discovering conserved RNA substructures from small RNA elements, like the kink-turn motif, to entire conserved domains. These substructures provide further information about their functional and evolutionary relations and can be helpful in regulatory activities.

6. Challenges

Although a long history of research is behind the motif discovery, there is still a long way to go. One reason arises from the indistinct definition of the motif itself. The definition of a motif is very generic and broad. In molecular biology, a motif is described as a subsequence of a biological sequence with a biological concept. Due to this vague definition, finding motifs becomes a problem of looking for something we do not fully know. The new findings in biology, biotechnology, and computational biology shed light on different aspects of motifs. This makes the process of motif discovery an ongoing problem that is way far from being fully understood and solved.

The second challenge arises from the interdisciplinary nature of motif discovery problem. The advancements in motif discovery usually come from other research communities, such as computer sciences and data engineering. New findings in pattern matching in the data mining area can directly impact the way motifs are discovered. On the other hand, the priorities and research problems are different in the two fields. So, sometimes there might be a problem that is very serious and challenging in motif discovery. However, because the problem is not interesting for computer scientists, it might remain intact for a while. The gap between the two communities is meant to be filled by interdisciplinary majors such as bioinformatics. The closer the researchers in different areas work together, the better the outcomes will be. We are still at the beginning of this way and the gap is yet to be filled.

Biological real-world shape of the motif is another challenge that sometimes makes it very hard to be discovered. The position of the motif might be unknown. The direction of the sequence is not fixed, and they might appear in entirely reverse directions. There might be gaps in the motifs. These challenges along with many other indeterminate aspects of motifs make it very hard to discover deterministically. In most cases, researchers talk about the probability of being motifs rather than deciding specifically whether a sequence is a motif or not.

Last but not least, tapping into the human computation power in solving motif discovery problems is another gap that needs investigations and experiments. The extraordinary capability of humans in solving complex problems, mainly when relying on a large crowd, is shown and proved in many areas such as image and text analysis, audio and video prescription, micro-tasking, and many more areas. Including human brain power in the computational algorithms and assessing its efficiency and impact is an exciting and challenging problem that needs to be invested in.

7. Conclusion and future directions

Finding motifs in RNA is a growing issue in computational biology. Motif discovery problems can be essential for the analysis and understanding of biological information. In this paper, we presented a taxonomy for motif discovery in RNA. This taxonomy aimed to simplify understanding, analyzing, and comparing existing motif discovery approaches. We identified three main characterizing aspects for motif discovery, explained these aspects, and analyzed some selected related works using the proposed framework and aspects. We also discussed some challenges in the area of motif discovery that need more investigation.

We believe that in the future, there will be far more collaborations between researchers from different domains to solve ever-increasing problems in the domain of sequence analysis, specifically motif discovery. Furthermore, the

inclusion of human intelligence and the wisdom of the crowd in solving biological problems is one of the promising directions that will attract more researchers from different fields such as computer science, social science, and biology together, to solve more complex biological problems.

8. Conflicts of interest

The authors declare that there is no conflict of interest.

9. References

[1] D. Wang and A. Farhana, "Biochemistry, RNA Structure," in *StatPearls [Internet]*: StatPearls Publishing, 2022.

[2] S. Di Carlo, G. Politano, A. Savino, and A. Benso, "A systematic analysis of a mi-RNA inter-pathway regulatory motif," *Journal of clinical bioinformatics,* vol. 3, no. 1, pp. 1-14, 2013.

[3] N. B. Leontis, A. Lescoute, and E. Westhof, "The building blocks and motifs of RNA architecture," *Current opinion in structural biology,* vol. 16, no. 3, pp. 279-287, 2006.

[4] N. B. Leontis and E. Westhof, "Analysis of RNA motifs," *Current opinion in structural biology,* vol. 13, no. 3, pp. 300-308, 2003.

[5] W. Saenger, *Principles of Nucleic Acid Structure*. Springer New York, NY, 1984, pp. XX, 556.

[6] N. B. Leontis and E. Westhof, "Geometric nomenclature and classification of RNA base pairs," *Rna,* vol. 7, no. 4, pp. 499-512, 2001.

[7] J. Pohar, D. Lainšček, A. Kunšek, M.-M. Cajnko, R. Jerala, and M. Benčina, "Phosphodiester backbone of the CpG motif within immunostimulatory oligodeoxynucleotides augments activation of Toll-like receptor 9," *Scientific reports,* vol. 7, no. 1, pp. 1-11, 2017.

[8] Y. Chen and G. Varani, "RNA structure," *eLS,* 2010.

[9] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath, "RNAMotif, an RNA secondary structure definition and search algorithm," *Nucleic acids research,* vol. 29, no. 22, pp. 4724-4735, 2001.

[10] F. Fassetti, G. Greco, and G. Terracina, "Mining loosely structured motifs from biological data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 20, no. 11, pp. 1472-1489, 2008.

[11] *Data Mining in Bioinformatics* (Advanced Information and Knowledge Processing). Springer-Verlag London, 2005.

[12] Z. Lu and H. Y. Chang, "Decoding the RNA structurome," *Current opinion in structural biology,* vol. 36, pp. 142-148, 2016.

[13] Y. S. Tsai, S. M. Gomez, and Z. Wang, "Prevalent RNA recognition motif duplication in the human genome," *RNA,* vol. 20, no. 5, pp. 702-712, 2014.

[14] A. Jolma *et al.*, "Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences," *Genome research,* vol. 30, no. 7, pp. 962-973, 2020.

[15] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook, "RNA structural motifs: building blocks of a modular biomolecule," *Quarterly reviews of biophysics,* vol. 38, no. 3, p. 221, 2005.

[16] G. J. Quigley and A. Rich, "Structural domains of transfer RNA molecules," *Science,* vol. 194, no. 4267, pp. 796-806, 1976.

[17] H.-C. Huang, U. Nagaswamy, and G. E. Fox, "The application of cluster analysis in the intercomparison of loop structures in RNA," *Rna,* vol. 11, no. 4, pp. 412-423, 2005.

[18] J. C. Lee, J. J. Cannone, and R. R. Gutell, "The lonepair triloop: a new motif in RNA structure," *Journal of molecular biology,* vol. 325, no. 1, pp. 65-83, 2003.

[19] A. Szewczak and P. Moore, "The sarcin/ricin loop, a modular RNA," *Journal of molecular biology,* vol. 247, no. 1, pp. 81-98, 1995.

[20] D. Klein, T. Schmeing, P. Moore, and T. Steitz, "The kink-turn: a new RNA secondary structure motif," *The EMBO journal,* vol. 20, no. 15, pp. 4214-4221, 2001.

[21] S. A. STROBEL, P. L. ADAMS, M. R. STAHLEY, and J. WANG, "RNA kink turns to the left and to the right," *Rna,* vol. 10, no. 12, pp. 1852-1854, 2004.

[22] S. Szep, J. Wang, and P. B. Moore, "The crystal structure of a 26-nucleotide RNA containing a hook-turn," *Rna,* vol. 9, no. 1, pp. 44-51, 2003.

[23] J. Sussman and S. Kim, "Absence of correlation between base-pair sequence and RNA conformation," *Science,* vol. 212, no. 4500, pp. 1275-1277, 1981.

[24] T. Hermann and D. J. Patel, "RNA bulges as architectural and recognition motifs," *Structure,* vol. 8, no. 3, pp. R47-R54, 2000.

[25] D. E. Draper, "A guide to ions and RNA structure," *Rna,* vol. 10, no. 3, pp. 335-343, 2004.

[26] R. T. Batey, R. P. Rambo, and J. A. Doudna, "Tertiary motifs in RNA structure and folding," *Angewandte Chemie International Edition,* vol. 38, no. 16, pp. 2326-2343, 1999.

[27] S. Kim *et al.*, "The general structure of transfer RNA molecules," *Proceedings of the National Academy of Sciences,* vol. 71, no. 12, pp. 4970-4974, 1974.

[28] J. H. Cate *et al.*, "Crystal structure of a group I ribozyme domain: principles of RNA packing," *Science,* vol. 273, no. 5282, pp. 1678-1685, 1996.

[29] K.-Y. Chang and I. Tinoco, "Characterization of a" kissing" hairpin complex derived from the human immunodeficiency virus genome," *Proceedings of the National Academy of Sciences,* vol. 91, no. 18, pp. 8705-8709, 1994.

[30] P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner, "SCOR: a structural classification of RNA database," *Nucleic acids research,* vol. 30, no. 1, pp. 392-394, 2002.

[31] E. Ennifar, P. Walter, B. Ehresmann, C. Ehresmann, and P. Dumas, "Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site," *Nature structural biology,* vol. 8, no. 12, pp. 1064-1068, 2001.

[32] P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz, "RNA tertiary interactions in the large ribosomal subunit: the A-minor motif," *Proceedings of the National Academy of Sciences,* vol. 98, no. 9, pp. 4899-4903, 2001.

[33] S. R. Holbrook, J. L. Sussman, R. W. Warrant, and S.-H. Kim, "Crystal structure of yeast phenylalanine transfer RNA: II. Structural features and functional implications," *Journal of molecular biology,* vol. 123,

no. 4, pp. 631-660, 1978.
[34]S. Holbrook and S.-H. Kim, "Intercalation conformations in single-and double-stranded nucleic acids," *International Journal of Biological Macromolecules,* vol. 1, no. 5, pp. 233-240, 1979.
[35]F. Van Batenburg, A. P. Gultyaev, and C. W. Pleij, "PseudoBase: structural information on RNA pseudoknots," *Nucleic acids research,* vol. 29, no. 1, pp. 194-195, 2001.
[36]L. X. Shen and I. Tinoco Jr, "The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus," *Journal of molecular biology,* vol. 247, no. 5, pp. 963-978, 1995.
[37]T. L. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Machine learning,* vol. 21, no. 1-2, pp. 51-80, 1995.
[38]T. L. Bailey *et al.*, "MEME SUITE: tools for motif discovery and searching," *Nucleic acids research,* vol. 37, no. suppl_2, pp. W202-W208, 2009.
[39]Z. Yao, Z. Weinberg, and W. L. Ruzzo, "CMfinder—a covariance model based RNA motif finding algorithm," *Bioinformatics,* vol. 22, no. 4, pp. 445-452, 2006.
[40]M. Rabani, M. Kertesz, and E. Segal, "Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes," *Proceedings of the National Academy of Sciences,* vol. 105, no. 39, pp. 14885-14890, 2008.
[41]M. Hiller, R. Pudimat, A. Busch, and R. Backofen, "Using RNA secondary structures to guide sequence motif finding towards single-stranded regions," *Nucleic acids research,* vol. 34, no. 17, pp. e117-e117, 2006.
[42]J. Han and M. Kamber, "Data Mining: Concepts and Techniques, 2nd editionMorgan Kaufmann Publishers," *San Francisco, CA, USA,* 2006.
[43]A. Achar and P. Sætrom, "RNA motif discovery: a computational overview," *Biology direct,* vol. 10, no. 1, p. 61, 2015.
[44]M. Quadrini, L. Tesei, and E. Merelli, "An algebraic language for RNA pseudoknots comparison," *BMC bioinformatics,* vol. 20, no. 4, pp. 1-18, 2019.
[45]M. Quadrini, L. Tesei, and E. Merelli, "ASPRAlign: a tool for the alignment of RNA secondary structures with arbitrary pseudoknots," *Bioinformatics,* vol. 36, no. 11, pp. 3578-3579, 2020.
[46]M. Quadrini, "Structural relation matching: an algorithm to identify structural patterns into RNAs and their interactions," *Journal of Integrative Bioinformatics,* 2021.
[47]G. Mauri and G. Pavesi, "Algorithms for pattern matching and discovery in RNA secondary structure," *Theoretical Computer Science,* vol. 335, no. 1, pp. 29-51, 2005.
[48]M. Anwar, T. Nguyen, and M. Turcotte, "Identification of consensus RNA secondary structures using suffix arrays," *BMC bioinformatics,* vol. 7, no. 1, p. 244, 2006.
[49]G. Badr, I. Al-Turaiki, M. Turcotte, and H. Mathkour, "IncMD: Incremental trie-based structural motif discovery algorithm," *Journal of bioinformatics and computational biology*, vol. 12, no. 05, p. 1450027, 2014.
[50]Y. Ji, X. Xu, and G. D. Stormo, "A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences," *Bioinformatics,* vol. 20, no. 10, pp. 1591-1602, 2004.
[51]M. Miladi *et al.*, "GraphClust2: annotation and discovery of structured RNAs with scalable and accessible integrative clustering," *GigaScience,* vol. 8, no. 12, p. giz150, 2019.
[52]M. Hamada, K. Tsuda, T. Kudo, T. Kin, and K. Asai, "Mining frequent stem patterns from unaligned RNA sequences," *Bioinformatics,* vol. 22, no. 20, pp. 2480-2487, 2006.
[53]D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," *SIAM journal on applied mathematics,* vol. 45, no. 5, pp. 810-825, 1985.
[54]J. Gorodkin, L. J. Heyer, and G. D. Stormo, "Finding the most significant common sequence and structure motifs in a set of RNA sequences," *Nucleic acids research,* vol. 25, no. 18, pp. 3724-3732, 1997.
[55]J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers: Original Research on Biomolecules,* vol. 29, no. 6-7, pp. 1105-1119, 1990.
[56]G. Badr, I. Al-Turaiki, and H. Mathkour, "Classification and assessment tools for structural motif discovery algorithms," *BMC bioinformatics,* vol. 14, no. S9, p. S4, 2013.
[57]J.-H. Chen, S.-Y. Le, and J. V. Maizel, "Prediction of common secondary structures of RNAs: a genetic algorithm approach," *Nucleic Acids Research,* vol. 28, no. 4, pp. 991-999, 2000.
[58]Y. J. Hu, "Prediction of consensus structural motifs in a family of coregulated RNA sequences," *Nucleic acids research,* vol. 30, no. 17, pp. 3886-3893, 2002.
[59]S. Michal, T. Ivry, O. Cohen, M. Sipper, and D. Barash, "Finding a common motif of RNA sequences using genetic programming: The GeRNAMo system," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 4, no. 4, pp. 596-610, 2007.
[60]K. L. Povl, L. Tommy, and M. Ulf, *Textbook of Drug Design and Discovery* Third edition ed. USA and Canada: Taylor & Francis, 2005.
[61]W. Yin *et al.*, "Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir," *Science,* 2020.
[62]S. Richter, H. Cao, and T. M. Rana, "Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1− Tat− TAR ternary complex formation," *Biochemistry,* vol. 41, no. 20, pp. 6391-6397, 2002.
[63]M. P. Robertson, H. Igel, R. Baertsch, D. Haussler, M. Ares Jr, and W. G. Scott, "The structure of a rigorously conserved RNA element within the SARS virus genome," *PLoS biology,* vol. 3, no. 1, p. e5, 2005.
[64]A. H. Aldhumani *et al.*, "RNA sequence and ligand binding alter conformational profile of SARS-CoV-2 stem loop II motif," *Biochemical and biophysical research communications,* vol. 545, pp. 75-80, 2021.
[65]A. Umuhire Juru, N. N. Patwardhan, and A. E. Hargrove, "Understanding the contributions of conformational changes, thermodynamics, and kinetics of RNA–small molecule interactions," *ACS chemical biology,* vol. 14, no. 5, pp. 824-838, 2019.

[66]S. Ramírez-Clavijo and G. Montoya-Ortíz, "Gene expression and regulation," in *Autoimmunity: From Bench to Bedside [Internet]*: El Rosario University Press, 2013.
[67]L. Nasalean, J. Stombaugh, C. Zirbel, and N. Leontis, "Non-protein coding RNAs," ed: Springer Berlin Heidelberg, 2009.
[68]K. A. Afonin *et al.*, "In vitro assembly of cubic RNA-based scaffolds designed in silico," *Nature nanotechnology*, vol. 5, no. 9, pp. 676-682, 2010.
[69]H. Saito and T. Inoue, "Synthetic biology with RNA motifs," *The international journal of biochemistry & cell biology*, vol. 41, no. 2, pp. 398-404, 2009.
[70]I. Severcan, C. Geary, A. Chworos, N. Voss, E. Jacovetty, and L. Jaeger, "A polyhedron made of tRNAs," *Nature chemistry*, vol. 2, no. 9, pp. 772-779, 2010.
[71]N. N. Qader and H. K. Al-Khafaji, "Motif discovery and data mining in bioinformatics," *Int. J. Comput. Technol*, vol. 13, no. 1, pp. 4082-4095, 2014.
[72]J. Xiong, "Essential Bioinformatics Cambridge University press," *Newyork. USA*, 2006.
[73]M. Vahed, M. Vahed, and L. X. Garmire, "BML: a versatile web server for bipartite motif discovery," *bioRxiv*, 2021.
[74]P. E. Compeau, P. A. Pevzner, and G. Tesler, "Why are de Bruijn graphs useful for genome assembly?," *Nature biotechnology*, vol. 29, no. 11, p. 987, 2011.
[75]T. Gao, J. Shu, and J. Cui, "A systematic approach to RNA-associated motif discovery," *BMC genomics*, vol. 19, no. 1, p. 146, 2018.
[76]M. Sarver, C. L. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis, "FR3D: finding local and composite recurrent structural motifs in RNA 3D structures," *Journal of mathematical biology*, vol. 56, no. 1-2, pp. 215-252, 2008.
[77]D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," *Proceedings of the National Academy of Sciences*, vol. 101, no. 19, pp. 7287-7292, 2004.
[78]G. Pavesi, P. Mereghetti, F. Zambelli, M. Stefani, G. Mauri, and G. Pesole, "MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W566-W570, 2006.
[79]G. Pavesi, G. Mauri, M. Stefani, and G. Pesole, "RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences," *Nucleic acids research*, vol. 32, no. 10, pp. 3258-3269, 2004.
[80]G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic acids research*, vol. 32, no. suppl_2, pp. W199-W203, 2004.
[81]M. Pietrosanto, M. Adinolfi, R. Casula, G. Ausiello, F. Ferrè, and M. Helmer-Citterich, "BEAM web server: a tool for structural RNA motif discovery," *Bioinformatics*, vol. 34, no. 6, pp. 1058-1060, 2018.
[82]A. Guarracino *et al.*, "BRIO: a web server for RNA sequence and structure motif scan," *Nucleic Acids Research*, 2021.
[83]E. Mattei, G. Ausiello, F. Ferre, and M. Helmer-Citterich, "A novel approach to represent and compare RNA secondary structures," *Nucleic acids research*, vol. 42, no. 10, pp. 6146-6157, 2014.
[84]M. Pietrosanto *et al.*, "Relative Information Gain: Shannon entropy-based measure of the relative structural conservation in RNA alignments," *NAR genomics and bioinformatics*, vol. 3, no. 1, p. lqab007, 2021.
[85]J. Ule, K. B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R. B. Darnell, "CLIP identifies Nova-regulated RNA networks in the brain," *Science*, vol. 302, no. 5648, pp. 1212-1215, 2003.
[86]J. König *et al.*, "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution," *Nature structural & molecular biology*, vol. 17, no. 7, pp. 909-915, 2010.
[87]M. Hafner *et al.*, "Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP," *Cell*, vol. 141, no. 1, pp. 129-141, 2010.
[88]D. Ray *et al.*, "Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins," *Nature biotechnology*, vol. 27, no. 7, pp. 667-670, 2009.
[89]A. Munteanu, N. Mukherjee, and U. Ohler, "SSMART: sequence-structure motif identification for RNA-binding proteins," *Bioinformatics*, vol. 34, no. 23, pp. 3990-3998, 2018.
[90]J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, "RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning," *Nature communications*, vol. 10, no. 1, pp. 1-13, 2019.
[91]W. Dawson, T. Takai, N. Ito, K. Shimizu, and G. Kawai, "A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped?," *Journal of Nucleic Acids Investigation*, vol. 5, no. 1, 2014.
[92]D. Maticzka, S. J. Lange, F. Costa, and R. Backofen, "GraphProt: modeling binding preferences of RNA-binding proteins," *Genome biology*, vol. 15, no. 1, pp. 1-18, 2014.
[93]E. Bindewald, T. Kluth, and B. A. Shapiro, "CyloFold: secondary structure prediction including pseudoknots," *Nucleic acids research*, vol. 38, no. suppl_2, pp. W368-W372, 2010.
[94]T. L. Bailey, "STREME: Accurate and versatile sequence motif discovery," *Biorxiv*, 2020.
[95]S. Engelen and F. Tahi, "Tfold: efficient in silico prediction of non-coding RNA secondary structures," *Nucleic acids research*, vol. 38, no. 7, pp. 2453-2466, 2010.
[96]X. Chen, N. S. Khan, and S. Zhang, "LocalSTAR3D: a local stack-based RNA 3D structural alignment tool," *Nucleic acids research*, vol. 48, no. 13, pp. e77-e77, 2020.